(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 813 184 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
17.12.1997 Bulletin 1997/51

(51) Int Cl.^{6}: G10L 5/04, G10L 3/02

(21) Application number: 97870079.7

(22) Date of filing: 29.05.1997

(84) Designated Contracting States:
BE CH DE DK ES FR GB IT LI NL SE

(30) Priority: 10.06.1996 BE 9600524

(71) Applicant: Faculté Polytechnique de Mons
7000 Mons (BE)

(72) Inventors:
- Dutoit, Thierry
  7332 Sirault (BE)
- Pagel, Vincent
  54000 Nancy (FR)
- Pierret, Nicolas
  7030 Saint-Symphorien (BE)

(74) Representative: Van Malderen, Joelle et al
Office Van Malderen,
Place Reine Fabiola 6/1
1083 Bruxelles (BE)

(54) **Method for audio synthesis**

(57) Method for audio signal synthesis from elementary audio waveforms stored in a dictionary characterized in that:

- the waveforms are perfectly periodic, and stored as one of their period.
- synthesis is obtained by overlap-adding of the waveforms obtained from time-domain multiplication of the periodic waveforms with a weighting window whose size is approximately two times the period of the signals to weight, and whose relative position inside of the period is fixed to any value identical for all the periods:

whereby the time shift between two successive waveforms obtained by weighting the original signals is set to the imposed fundamental frequency of the signal to synthesize.

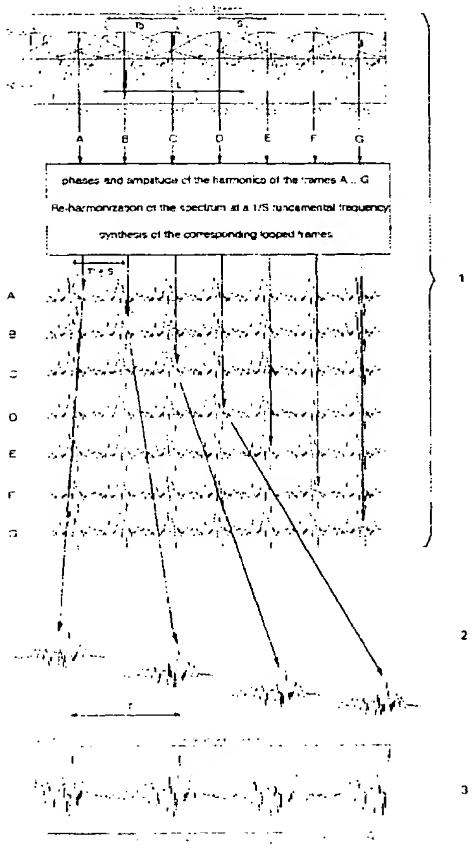
A
B
C
D
E
F
G
phases and amplitude of the harmonics of the frames A .. G
Re-harmonization of the spectrum at a 1/S fundamental frequency
synthesis of the corresponding looped frames
T = S
A
B
C
D
E
F
G
1
2
3

FIG. 1

EP 0 813 184 A1

Printed by Jouve, 75001 PARIS (FR)

Description

The invention described herein relates to a method of synthesis of audio sounds. To simplify the description, focus is mainly made on vocal sounds, keeping in mind that the invention can be applied to the field of music synthesis as well.

Background of the invention

In the framework of the so-called "concatenative" synthesis techniques which are increasingly used, synthetic speech is produced from a database of speech segments. Segments may be diphones, for example, which begin from the middle of the stationary part of a phone (the phone being the acoustic realization of a phoneme) and end in the middle of the stationary part of the next phone. French, for instance, is composed of 36 phonemes, which corresponds to approximately 1240 diphones (as a matter of fact some combination of phonemes are impossible). Other types of segments can be used, like triphones, polyphones, half-syllables, etc. Concatenative synthesis techniques produce any sequence of phonemes by concatenating the appropriate segments. The segments are themselves obtained from the segmentation of a speech corpus read by a human speaker.

Two problems must be solved during the concatenation process in order to get a speech signal comparable to human speech.

The first problem arises from the disparities of the phonemic contexts from which the segments were extracted, which generally results in some spectral envelope mismatch at both ends of the segments to be concatenated. As a result, a mere concatenation of segments leads to sharp transitions between units, and to less fluid speech.

The second problem is to control the prosody of synthetic speech, i.e. its rhythm (phoneme and pause lengths) and its fundamental frequency (the vibration frequency of the vocal folds). The point is that the segments recorded in the corpus have their own prosody that does not necessarily correspond to the prosody imposed at synthesis time.

Hence there is a need to find a means of controlling prosodic parameters and of producing smooth transitions between segments, without affecting the naturalness of speech segments.

One distinguishes two families of methods to solve such problems: the ones that implement a spectral model of the vocal tract, and the ones that modify the segment waveforms directly in the time domain.

In the first category of methods, transitions between concatenated segments are smoothed out by computing the difference between the spectral envelopes on both sides of the concatenation point, and propagating this difference in the spectral domain on both segments. The way it controls the pitch and the duration of segments depends on the particular model used for spectral envelope estimation. All these methods require a high computational load at synthesis time, which prevents them from being implemented in real time on low-cost processors.

On the contrary the second family of synthesis methods aims to produce concatenation and prosody modification directly in the time domain with very limited computational load. All of them take advantage of the so-called "Poisson's Sum Theorem", well known among signal processing specialists which demonstrates that it is possible to build from any finite waveform with a given spectral envelope an infinite waveform with the same spectral envelope for an arbitrarily chosen (and constant) pitch. This theorem can be applied to the modification of the fundamental frequency of speech signals. Provided the spectrum of the elementary waveforms is close to the spectral envelope of the signal one wishes to modify, pitch can be imposed by setting the shift between elementary waveforms to the targeted pitch period, and by adding the resulting overlapping waveforms. In this second family, synthesis methods mainly differ in the way they derive elementary waveforms from the pre-recorded segments. However, in order to produce high-quality synthetic speech, the overlapping elementary waveforms they use must have a duration of at least twice the fundamental frequency of the original segments. Two classes of techniques in this second family of synthesis methods will be described hereafter.

The first class refers to methods hereafter referred to as 'PSOLA' methods (Pitch Synchronous Overlap Add), characterized by the direct extraction of waveforms from continuous audio signals. The audio signals used are either identical to the original signals (the segments), or obtained after some transformation of these original signals. Elementary waveforms are extracted from the audio signals by multiplying the signals with finite-duration weighting windows positioned synchronously with the fundamental frequency of the original signal. Since the size of the elementary waveforms must be at least twice the original period, and given that there is one waveform for each period of the original signal, the same speech samples are used in several successive waveforms: the weighting windows overlap in the audio signals.

Examples of such PSOLA methods are those defined in documents EP-0363233, US-5479564, EP-0706170. A specific example is also the MBR-PSOLA method as published by T. Dutoit and H. Leich, in Speech Communication, Elsevier Publisher, November 1993, Vol. 13, N° 3-4, 1993. The method described in document US-5479564 suggests a means of modifying the frequency of an audio signal with constant fundamental frequency by overlap-adding short-term signals extracted from this signal. The length of the weighting windows used to obtain the short-term signals is approximately equal to two times the period of the audio signal and their position within the period can be set to any value (provided the time shift between successive windows is equal to the period of the audio signal). Document US-5479564 also describes a means of interpolating waveforms between segments to concatenate. so as to smooth out discontinuities. This is achieved by modifying the periods corresponding to the end of the first segment and to the beginning of the second segment. in such a way as to propagate the difference between the last period of the first segment and the first period of the second segment.

The second class of techniques, hereafter referred to as 'analytic', is based on a time-domain modification of waveforms that do not share, even partially, their samples. The synthesis step still uses shifting and overlap-adding of the weighted waveforms carrying the spectral envelope information. These waveforms are no longer extracted from a continuous speech signal by means of overlapping weighting windows. Examples of these techniques are those defined in documents US-5369730 and GB-2261350. as well as by T. Yazu. K. Yamada. "The speech synthesis system for an unlimited Japanese vocabulary", in proceedings IEEE ICASSP 1986. Tokyo. pp. 2019-2022.

In all these 'analytic' techniques. elementary waveforms are impulse responses of the vocal tract computed from evenly spaced speech signal frames. and resynthesized via a spectral model. The present invention falls in this class of methods.

An advantage of analytic methods over PSOLA methods is that the waveforms they use result from a true spectral model of the vocal tract. Therefore, they can intrinsically model the instantaneous spectral envelope information with more accuracy and precision than PSOLA techniques, which simply weight a time-domain signal with a weighting window. Moreover, it is possible with analytic methods to separate the periodic (voiced) and aperiodic (unvoiced) components of each waveform, and modify their balance during the resynthesis step in order to modify the speech quality (soft, harsh, whispered, etc).

In practice, this advantage is counterbalanced by an increase of the size of the resynthesized segment database (typically a factor 2 since the successive waveforms do not share any samples while their duration still has to be equal to at least two times that of the pitch period of the audio signal). The method described by MM. Yazu and Yamada precisely aims at reducing the number of samples to be stored, by resynthesizing impulse responses in which the phases of the spectral envelope are set to zero. Only half of the waveform needs to be stored in this case. since phase zeroing results in perfectly symmetrical waveforms. The main drawback of this method is that it greatly affects the naturalness of the synthetic speech. It is well known. indeed, that performing important phase distortions have a strong effect on speech quality.

Aim of the invention

The present invention aims to suggest a method for audio synthesis that avoids the drawbacks presented in the state of the art and which requires limited storage for the waveforms while avoiding important distortions of the natural phase of acoustic signals.

Main characteristic elements of the invention

The present invention relates to a method for audio synthesis from waveforms stored in a dictionary characterized by the following points:

- the waveforms are infinite and perfectly periodic, and are stored as one of their period, itself represented as a sequence of sound samples of a priori of any length:
- Synthesis is carried out by overlapping and adding the waveforms multiplied by a weighting window whose length is approximately two times the period of the original waveform, and whose position relatively to the waveform can be set to any fixed value:

The time shift between two successive weighted signals obtained by weighting the original waveforms is equal to the fundamental period requested for the synthetic signal, whose value is imposed. This value may be lower or greater than that of the original waveforms.

The method according to the present invention. basically differs from any other `analytic' method by the fact that the elementary waveforms used are not impulse responses of the vocal tract, but infinite periodic signals. multiplied by a weighting window to keep their length finite. and carrying the same spectral envelope as the original audio signals. A spectral model (hybrid harmonic/stochastic model, for instance. although the invention is not exclusively related to any particular spectral model) is used for resynthesis in order to get periodic waveforms (instead of the symmetric impulse responses of MM. Yazu and Yamada) carrying instantaneous spectral envelope information. Because of the periodicity of the elementary waveforms produced, only the first period need to be stored. The sound quality obtained by this method is incomparably superior to the one of MM. Yazu and Yamada. since the computation of the periodic waveforms do not impose phase constraints on the spectral envelopes, thereby avoiding the related quality degradation.

The periods that need to be stored are obtained by spectral analysis of a dictionary of audio segments (e. g. diphones in the case of speech synthesis). Spectral analysis produces spectral envelope estimates throughout each segment. Harmonic phases and amplitudes are then computed from the spectral envelope and the target period (i.e. the spectral envelope is sampled with the targeted fundamental frequency).

The length of each resynthesis period can advantageously be chosen equal for all the periods of all the segments. In this particular case. classical techniques for waveform compression (e.g. ADPCM) allow very high compression ratios (about 8) with very limited computational cost for decoding. The remarkable efficiency of such techniques on the waveforms obtained mainly originates from the fact that:

- all the periods stored in the segment database have the same length, which leads to a very efficient period to period differential coding scheme:
- the use of a spectral model for spectral envelope estimation allows the separation of harmonic and stochastic components of the waveforms. When the energy of the stochastic component is low enough compared to that of the harmonic component, it may be completely eliminated, in which case only the harmonic component is resynthesized. This results in waveforms that are more pure, noiseless, and exhibit more regularity than the original signal. which additionally enhances the efficiency of ADPCM coding techniques.

To further enhance the efficiency of coding techniques. the phases of the lower-order (i.e., lower frequency) harmonics of each stored period may be fixed (one phase value fixed for each harmonic of the database) for the resynthesis step. The frequency band where this setting is acceptable ranges from 0 to approximately 3 kHz. In this case. the resynthesis operation results in a sequence of periods with constant length, in which the time-domain difference between two successive periods is mainly due to spectral envelope differences. Since the spectral envelope of audio signals generally changes slowly with time, given the inertia of the physical mechanisms that produce them. the shape of the periods obtained in this way also evolve slowly. This. in turn, is particularly efficient when it comes to coding signals on the basis of period to period differences.

Independently of its use for segment coding, the idea of imposing a set of fixed values for the phases of the lower frequency harmonics leads to the implementation of a temporal smoothing technique between successive segments. to attenuate spectral mismatch between periods. The temporal difference between the last period of the first segment and the first period of the second segment is computed. and smoothly propagated on both sides of the concatenation point with a weighting coefficient continuously varying from -0.5 to 0.5 (depending on which side of the concatenation point is processed).

It should be noted that although the efficient coding properties and smoothing capabilities mentioned above were already available in the MBR-PSOLA technique as described in the state of the art. their effect is drastically reinforced in the present invention as opposed to the waveforms used by MBR-PSOLA. the periods used here do not share any of their samples. allowing a perfect separation between harmonically purified waveforms, and waveforms that are mainly stochastic. . .

Finally. the present invention still makes it possible to increase the quality of the synthesized audio signal by associating, with each resynthesized segment (or 'base segment'). a set of replacement segments similar but not identical to the base segment. Each base segment is processed in the same way as the corresponding base segment. and a sequence of periods is resynthesized. For each replacement segment. for instance. one can keep two periods corresponding respectively to the beginning and the end of the replacement segment at synthesis time. When two segments are about to be concatenated. it is then possible to modify the periods of the first base segment so as to propagate, on the last periods of this segment. the difference between the last period of the base segment and the last period of one of its replacement segments. Similarly, it is possible to modify the periods of the second base segment so as to propagate, on the first periods of this segment, the difference between the first period of the base segment and the first period of one of its replacement segments. The propagation of these differences is simply performed by multiplying the differences by a weighting coefficient continuously varying from 1 to 0 (from period to period) and addinq the weiqhted differences to the periods of the base segments.

Such a modification of the time-domain periods of a base segment so as to make it sound like one of its replacement segments can be advantageously used to produce free variants to a base sound, thereby avoiding the monotony resulting from the repeated use of a base sound. It can also be put to use for the production of linguistically motivated sound variants (e.g., stressed/unstressed vowels, tense/soft voice, etc.)

The fundamental difference between the method described in the state of the art, which according to our classification is a 'PSOLA' method, and the method of the present invention originates in the particular way of deriving the periods used. As opposed to the waveforms extracted from a continuous signal as proposed in the state of the art, the waveforms used in the present invention do not share any of their samples (hence, they do not overlap). It therefore benefits from the typical advantages of other analytic methods:

- very efficient coding techniques which account for the fact that:
    - periods can be harmonically purified by completely eliminating their stochastic component:
    - when resynthesizing periods, the phase of low-frequency harmonics can bet set constant (i.e.. one fixed value for each harmonic throughout the segment database)
- Ability to produce sound variants by interpolating

between base and replacement segments. For each base segment, for instance, two additional periods are stored, corresponding to the beginning and end of the segment and taken from a replacement segment. This enables the synthesis of more natural sounding voices.

Brief description of the drawings

The method according to the present invention shall be more precisely described by comparing it with the following state-of-the-art methods:

Figure 1 illustrates the different steps of speech synthesis according to a PSOLA method,

Figure 2 describes the different steps of speech synthesis according to the method proposed by MM. Yazu and Yamada,

Figure 3 describes the different steps of speech synthesis in accordance to the present invention.

Description of a preferred embodiment of the invention

Figure 1 shows a classical representation of a PSOLA method characterized by the following steps:

1. At least on the voiced parts of speech segments, an analysis is performed by weighting speech with a window approximately centered on the beginning of each impulse response of the vocal tract excited by the vocal folds. The weighting window has a shape that decreases down to zero at its edges, and its length is at least approximately two times the fundamental period of the original speech, or two times the fundamental period of the speech to be synthesized.
2. The signals that result from the weighting operation are shifted from each other, the shift being adjusted to the fundamental period of the speech to be synthesized, lower or greater than the original one, following the prosodic information related to the fundamental period at synthesis time.
3. Synthetic speech is obtained by summing these shifted signals.

Figure 2 shows the method described by MM. Yazu and Yamada according to the state of the art which implements 3 steps:

1. The original speech is cut out every fixed frame period (hence, not pitch synchronously), and the spectrum of each frame is computed by cepstral analysis. Phase components are set to zero, so that only spectral amplitudes are retained. A symmetric waveform is then obtained for each initial frame by inverse FFT. This symmetric waveform is weighted with a fixed length window that decreases to almost zero at its borders.
2. The signals that result from the weighting operation are shifted from each other, the shift being adjusted to the fundamental period of the speech to be synthesized, lower or greater than the original one, following the prosodic information related to the fundamental period at synthesis time.
3. Synthetic speech is obtained by summing these shifted signals.

In this last technique, steps 1 and 2 are often realized once for all, which makes the difference between analytic methods and those based on a spectral model of the vocal tract. The processed waveforms are stored in a database that centralizes, in a purely temporal format, all the information related to the evolution of the spectral envelope of the speech segments.

Concerning the preferred implementation of the invention herein described, figure 3 describes the following steps:

1. Analysis frames are assigned a fixed length and shift (denoted by S). Instead of estimating the spectral envelope of each analysis frame by cepstral analysis and computing its inverse FFT (as done by MM. Yazu and Yamada), the analysis algorithm of the powerful MBE (Multi-Band Excited) model is used, which computes the frequency, amplitude, and phase of each harmonic of the analysis frame. The spectral envelope is then derived for each frame and modify the frequencies and amplitudes of harmonics without changing this envelope, so as to obtain a fixed fundamental frequency equal to the analysis shift, S (i.e., the spectrum is "re-harmonized" in the frequency domain). Phases of the lower harmonics are set to a set of fixed values (i.e., a value chosen once for all for a given harmonic number). Time-domain waveforms are then obtained from harmonics by computing a sum of sinusoids, the frequencies, amplitudes, and phases are set equal to those of harmonics. As opposed to the invention of MM. Yazu and Yamada, the waveforms are not symmetrical, as phases have not been set to zero (there was no other choice in the previous method) . Furthermore, the precise waveforms obtained are not imposed by the algorithm, as they strongly depend on the fixed phase values imposed before resynthesis. Instead of storing the complete waveform in a segment database, one period of the waveform is only kept, since it is perfectly periodic by construction (sum of harmonics). This peridd can be unfolded to obtain the corresponding infinite waveform as required for the next step.
2. On the voiced parts of speech seqments, an analysis is performed by weighting the aforementioned re-synthesized waveform (obtained by looping one of its periods computed as a sum of harmonics) with

a window with fixed length. The weighting window has a shape that decreases down to zero at its edges. and its length is exactly two times the value of S. and therefore also two times the fundamental period the re-synthesized speech obtained in step 1 One such window is taken from each infinite waveform derived in step 1.
3. The signals that result from the weighting operation are overlapped and shifted from each other, the shift being adjusted to the fundamental period of the speech to be synthesized, lower or greater than S. following the prosodic information related to the fundamental period at synthesis time. Synthetic speech is obtained by summing these shifted signals.

The invention makes it possible to smooth out spectral discontinuities in the time domain due to the fixed set of phases applied to the periods during the resynthesis step for lower-order harmonics, since an interpolation between two such periods in the time-domain is then equivalent to an interpolation in the frequency domain.

Claims

1. Method for audio synthesis from waveforms stored in a dictionary, characterized in that the following steps are performed:

   - the waveforms are infinite and perfectly periodic, and are stored as one of their period, itself represented as a sequence of sound samples of a priori any length:
   - a synthesis is carried out by overlapping and adding the waveforms multiplied by a weighting window whose length is approximately two times the period of the original waveform, and whose position relatively to the waveform can be set to any fixed value:

   whereby the time shift between two successive weighted signals obtained by weighting the original waveforms is equal to the fundamental period requested for the synthetic signal. whose value is imposed.

2. Method for audio synthesis according to claim 1 characterized in that the fundamental period of the synthetic signal is greater or lower than the original period in the dictionary.

3. Method for audio synthesis according to claim 1 or 2 characterized in that the lengths of the periods stored in the dictionary are all identical

4. Method for audio synthesis according to claim 3 characterized in that the phases of the lower-frequency harmonics (typically from 0 to 3 kHz) of the stored periodic waveforms have a fixed value per harmonic throughout the dictionary.

5. Method for audio synthesis according to any of the preceding claims, characterized in that the stored waveforms are obtained from the spectral analysis of a dictionary of audio signal segments such as diphones in the case of speech synthesis whereby a spectral analysis provides at regular time intervals an estimate of the instantaneous spectral envelope in each segment from which the waveforms are computed.

6. Method for audio synthesis according to claim 5, characterized in that when concatenating two segments, the last periods of the first segment and the first period of the second segment are modified to smooth out the time-domain difference measured between the last period of the first segment and the first period of the second segment, this time-domain difference being added to each modified period with a weighting coefficient varying between -0.5 and 0.5 depending on the position of the modified period with respect to the concatenation point.

7. Method for audio synthesis according to claim 6, characterized in that for each base segment, replacement segments are stored whereby at synthesis time, when two segments are about to be concatenated, the periods of the first base segment are modified so as to propagate, on the last periods of this segment, the difference between the last period of the base segment and the last period of one of its replacement segments and whereby the periods of the second base segment are modified so as to propagate, on the first periods of this segment. the difference between the first period of the base segment and the first period of one of its replacement segments, the propagation of these differences being performed by multiplying the measured differences by a weighting coefficient continuously varying from 1 to 0 (from period to period) and adding the weighted differences to the periods of the base segments.

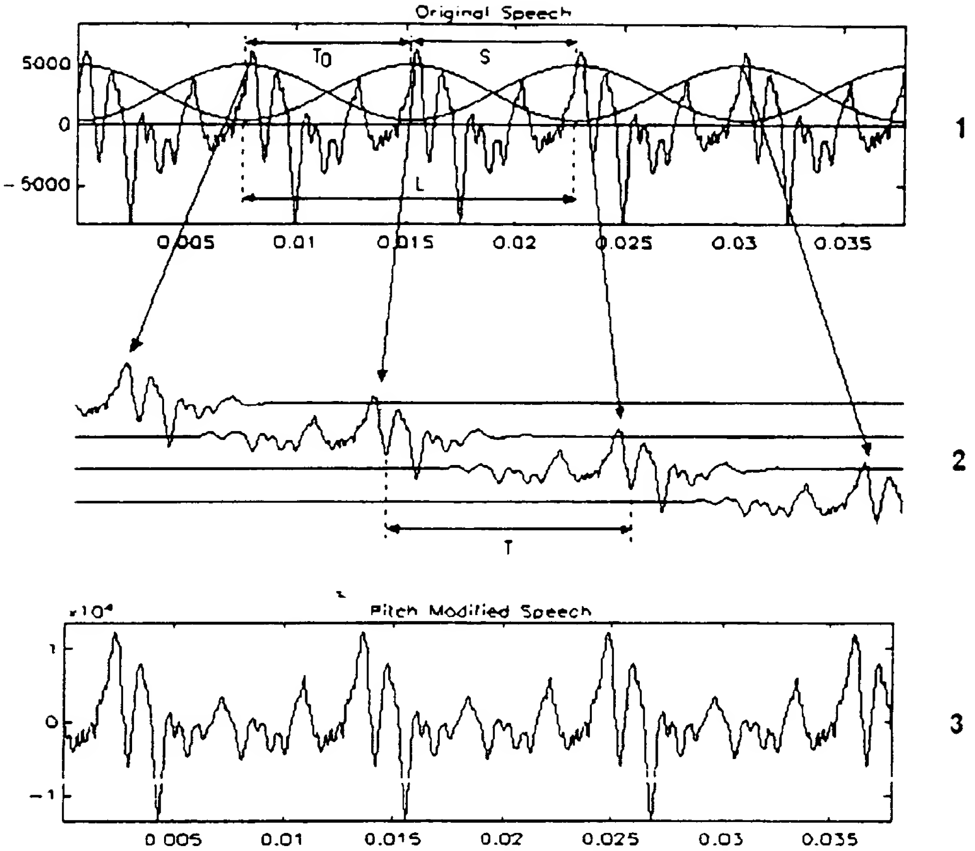
Original Speech
T0
S
5000
0
-5000
L
0.005
0.01
0.015
0.02
0.025
0.03
0.035
1
2
T
x10⁴
Pitch Modified Speech
1
0
-1
0.005
0.01
0.015
0.02
0.025
0.03
0.035
3

Fig. 1

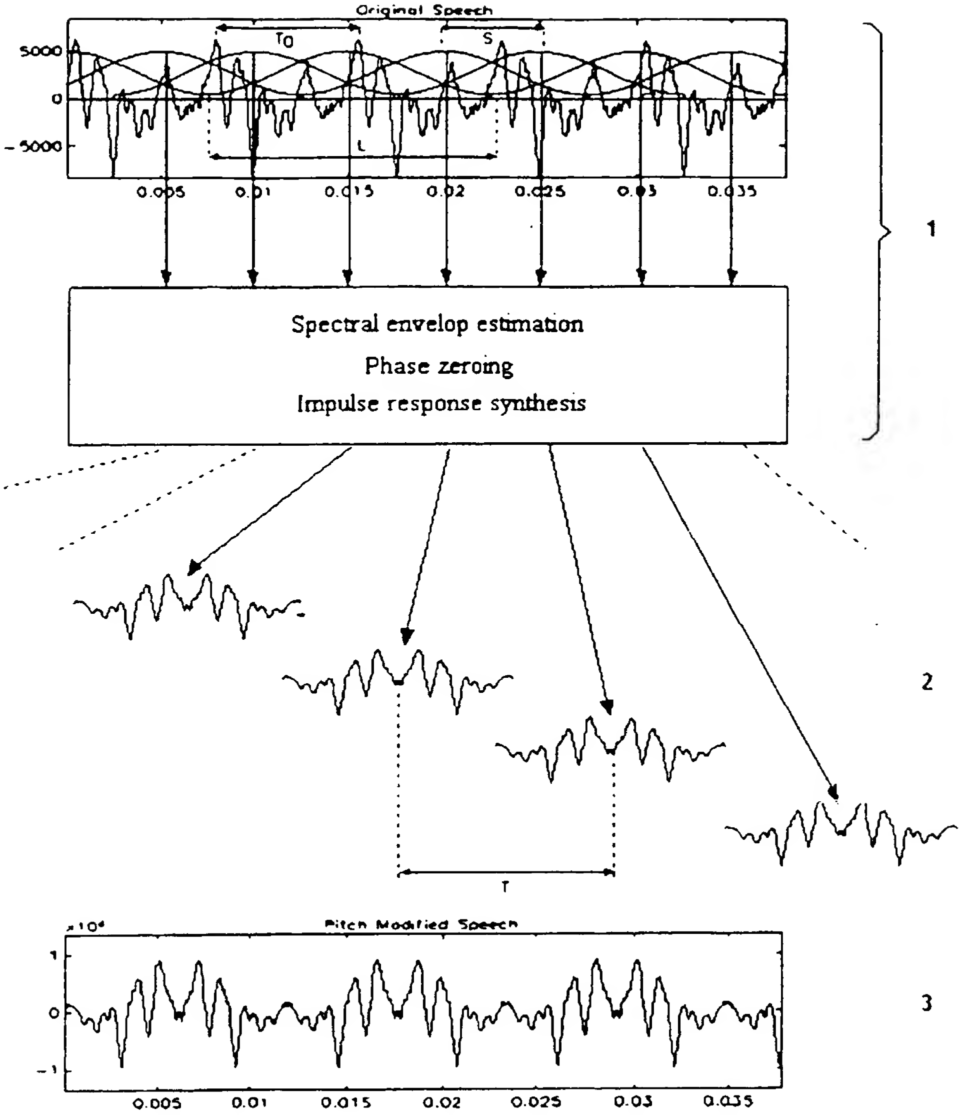
Original Speech
T0
S
L
5000
0
-5000
0.005
0.01
0.015
0.02
0.025
0.03
0.035
Spectral envelop estimation
Phase zeroing
Impulse response synthesis
1
2
T
3
x10⁴
Pitch Modified Speech
1
0
-1
0.005
0.01
0.015
0.02
0.025
0.03
0.035

Fig. 2

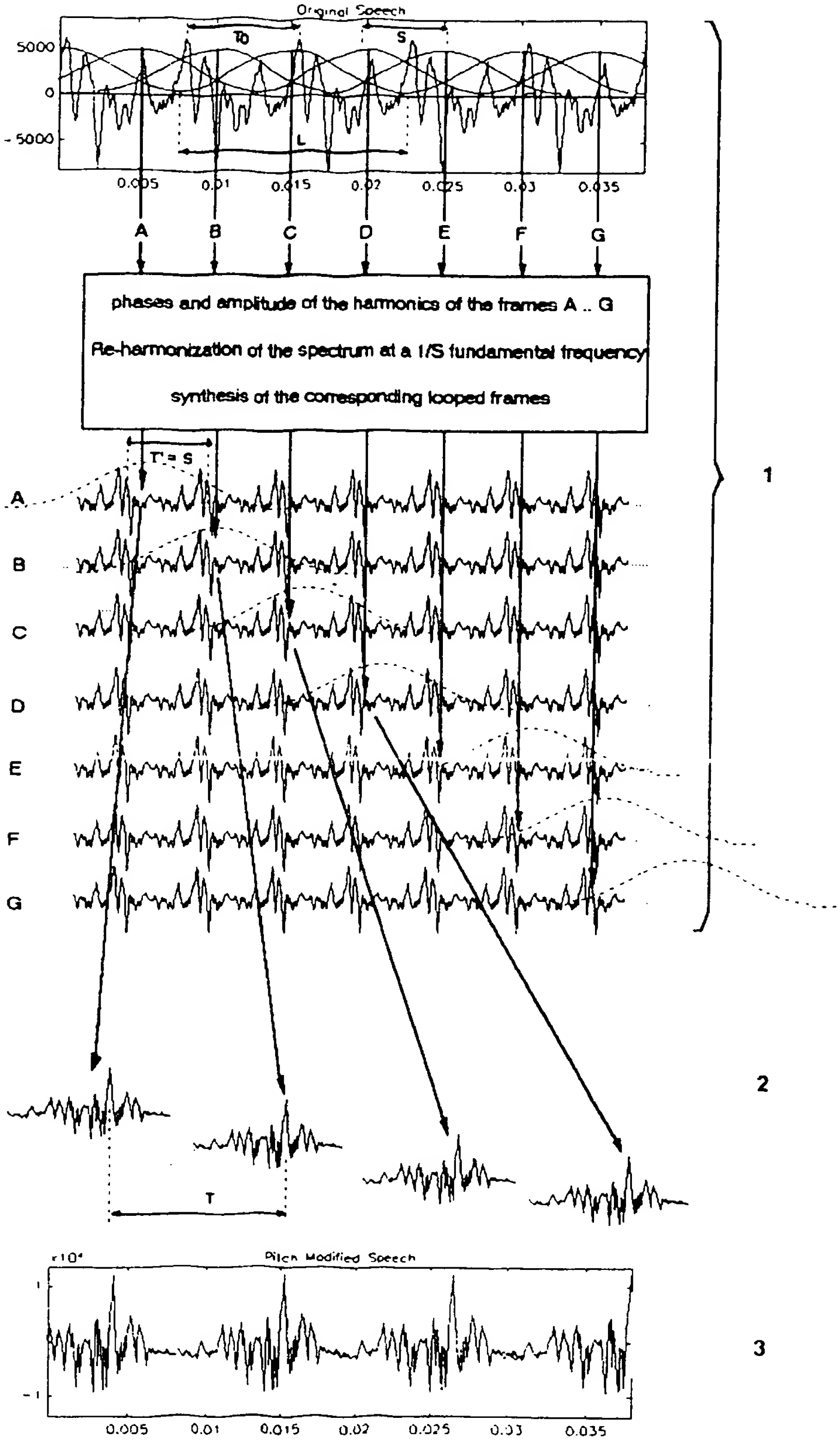
Original Speech
T0
S
5000
0
- 5000
L
0.005
0.01
0.015
0.02
0.025
0.03
0.035
A
B
C
D
E
F
G
phases and amplitude of the harmonics of the frames A .. G
Re-harmonization of the spectrum at a 1/S fundamental frequency
synthesis of the corresponding looped frames
T' = S
A
B
C
D
E
F
G
1
2
T
x10⁴
Pitch Modified Speech
1
-1
0.005
0.01
0.015
0.02
0.025
0.03
0.035
3

Fig. 3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 97 87 0079

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
X	EP 0 527 527 A (PHILIPS) 17 February 1993 * column 2, line 51 - column 4, line 17; figure 1 * ---	1-3	G10L5/04 G10L3/02
X	WO 90 03027 A (FRANCE ETAT) 22 March 1990 * page 4, line 10 - page 6, line 34 * ---	1-3	
A	COX ET AL.: "Real-time implementation of time-domain harmonic scaling of speech for rate modification and coding" IEEE JOURNAL OF SOLID-STATE CIRCUITS, vol. SC-18, no. 1, February 1983, pages 10-24, XP002026412 * page 10, right-hand column - page 11, left-hand column * ---	5	
A	VERHELST ET AL.: "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech" INTERNATIONAL CONFERENCE ON ACOUSTICS, , SPEECH AND SIGNAL PROCESSING 1993, vol. 2, 27 - 30 April 1993, MINNEAPOLIS, MN, US, pages 554-557, XP000427849 * page 555, left-hand column, line 25 - page 555, right-hand column, line 15 * ---	6	TECHNICAL FIELDS SEARCHED (Int.Cl.6) G10L
D,A	YAZU ET AL.: "The speech synthesis system for an unlimited Japanese vocabulary" INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING 1986, vol. 3, 7 - 11 April 1986, TOKYO, JP, pages 2019-2022, XP000567953 -----	1	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	17 September 1997	Lange, J

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document